SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB NO. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE December 1996	3. REPORT TYPE AND DATES COVERED Technical

4. TITLE AND SUBTITLE Canonical Coordinates for Graphical Representation of Multivariate Data	5. FUNDING NUMBERS DAAH04-96-1-0082
6. AUTHOR(S) C.R. Rao	
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES) Center for Multivariate Analysis 417 Thomas Building Department of Statistics Penn State University University Park, PA 16802	8. PERFORMING ORGANIZATION REPORT NUMBER 96-12
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	10. SPONSORING / MONITORING AGENCY REPORT NUMBER ARO 35518.6-MA

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited.	12 b. DISTRIBUTION CODE

19970210 231

13. ABSTRACT *(Maximum 200 words)*

A general theory is developed for representing population profiles characterized by multiple measurements in a low dimensional Euclidean space. The basic inputs of the problem are a matrix of distances between population profiles and the weights to be attached to different population profiles. The derived coordinates in the reduced space are called canonical coordinates. A well known method for representing row or column profiles in a contingency table using a chisquare type distance between profiles is the correspondence analysis. It is suggested that a similar analysis based on Hellinger distance between profiles has some advantages and is better suited for studying the configuration of profiles.

An asymmetric biplot technique which is useful in interpreting differences in row (column) profiles in terms of column (row) categories is developed.

14. SUBJECT TERMS Biplots, Canonical coordinates, Chisquare distance, Correspondence analysis, Hellinger distance, Principal component analysis	15. NUMBER IF PAGES 20
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL

CANONICAL COORDINATES FOR GRAPHICAL REPRESENTATION OF MULTIVARIATE DATA

C.Radhakrishna Rao

Technical Report 96-12

December 1996

Center for Multivariate Analysis
417 Thomas Building
Penn State University
University Park, PA 16802

Research sponsored by the Army Research Office under Grant DAAHO4-96-1-0082. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

Typeset by $\mathcal{AMS}$-TEX

CANONICAL COORDINATES FOR GRAPHICAL REPRESENTATION OF MULTIVARIATE DATA

by

C. Radhakrishna Rao

Department of Statistics
Pennsylvania State University
University Park, PA 16802

ABSTRACT

A general theory is developed for representing population profiles characterized by multiple measurements in a low dimensional Euclidean space. The basic inputs of the problem are a matrix of distances between population profiles and the weights to be attached to different population profiles. The derived coordinates in the reduced space are called canonical coordinates. A well known method for representing row or column profiles in a contingency table using a chisquare type distance between profiles is the correspondence analysis. It is suggested that a similar analysis based on Hellinger distance between profiles has some advantages and is better suited for studying the configuration of profiles.

An asymmetric biplot technique which is useful in interpreting differences in row (column) profiles in terms of column (row) categories is developed.

AMS Classification index: 62H30, 62H17.

Key Words and Phrases: Biplots, Canonical coordinates, Chisquare distance, Correspondence analysis, Hellinger distance, Principal component analysis.

1. INTRODUCTION

The problem we consider may be stated as follows. Let $X = (X_1 : \ldots : X_m)$ be a $p \times m$ data matrix with the i-th column X_i representing p measurements made on the i-th unit (population or individual). The vector X_i is called the i-th unit profile (UP). The UP's can be represented as points in an appropriate p-dimensional space R^p. The object is to construct a $k \times m$ matrix

$$Y = (Y_1 : \ldots : Y_m) \tag{1.1}$$

with $k < p$, for representing the UP's in a k-dimensional Euclidean space E^k ($Y_i \in E^k$ as the coordinate vector of the i-th unit) with the usual definition of distance between points in such a way that the configuration of points in R^p is maintained to the largest possible extent in E^k. To make the problem more precise, we need to specify the configuration of points relevant to a given problem and also a method of comparing the configurations. A natural way of specifying the configuration of points in a space is through a matrix of distances (or dissimilarities) between points. Let $\Delta = (\delta_{ij})$ be the $m \times m$ matrix of dissimilarities between the points $(X_1, \ldots, X_m)$ in R^p, and $D = (d_{ij})$ be the corresponding $m \times m$ matrix of Euclidean distances between the points $(Y_1, \ldots, Y_m)$ in E^k. We have to determine the transformation $X \rightarrow Y$ in such a way that $\Delta - D$ is close to the null matrix or, at least d_{ij}'s bear a monotone relationship with δ_{ij} to the extent possible. We consider some specific problems associated with continuous and discrete measurements.

2. CANONICAL COORDINATES

Let X be as defined in Section 1 with X_i as the column vector representing the p measurements taken on the i-th unit. Let R^p be a vector space with a specified inner product and the associated norm

$$\begin{aligned}(x, y) &= x'My, x, y \in R^p \\ \|x\| &= (x', x)^{1/2}, x \in R^p.\end{aligned} \tag{2.1}$$

where M is a positive definite matrix. Then the configuration of the m points $X_1, \ldots, X_m$ can be specified by the matrix $C = (c_{ij})$ where

$$c_{ij} = (X_i - \xi)'M(X_i - \xi) \tag{2.2}$$

and $\xi = (\xi_1, \ldots, \xi_p)'$ is a fixed point in R^p. We call such a vector space as Mahalanobis or M-space. Let us also associate with X_i a weight $w_i \geq 0$ such that $w_1 + \ldots + w_m = 1$, and define $W = \text{diag}(w_1, \ldots, w_m)$, i.e., a diagonal matrix with the elements $w_1, \ldots, w_m$ in the diagonal. The corresponding configuration with the origin as the reference point in the reduced space E^k is $Y'Y$ where Y is of order $k \times m$ with the i-th column representing the i-th population. We choose Y such that

$$\|C - YY'\| \tag{2.3}$$

is a minimum for a suitably chosen norm. We consider a (W, W)-invariant norm satisfying the condition

$$\|B(C - YY')B\| = \|C - YY'\| \;\forall B \ni B'WB = B \tag{2.4}$$

in which case the closed form solution to the problem is contained in the following theorem.

Theorem: Consider the s.v.d. (singular value decomposition)

$$M^{1/2}(X - \xi 1')W^{1/2} = \lambda_1 U_1 V_1' + \ldots + \lambda_p U_p V_p' \tag{2.5}$$

with singular values $\lambda_1 \geq \ldots \geq \lambda_p$, where $M^{1/2}$ and $W^{1/2}$ are symmetric square roots of M and W respectively. Then the choice of

$$Y = \begin{pmatrix} \lambda_1 V_1' W^{-1/2} \\ \vdots \\ \lambda_k V_k' W^{-1/2} \end{pmatrix} \tag{2.6}$$

minimizes (2.3) for any (W, W)-invariant norm.

The matrix Y in (2.6) is of order $k \times m$ and the i-th column gives the coordinates of the i-th unit in the reduced space E^k. We call these coordinates as the canonical coordinates. We note that a further minimization of (2.3) with respect to ξ leads to the choice

$$\hat{\xi} = XW1. \tag{2.7}$$

A detailed proof of the above theorem can be found in Rao (1964, 1979, 1980, 1985).)

The solution Y given in (2.6) can be expressed as the transformation

$$Y = G'(X - \xi 1') \tag{2.8}$$

where $G = M^{1/2}(U_1 : \ldots : U_k)$, which is of order $p \times k$, and U_i is as defined in the s.v.d. (2.5). It is seen that the columns of G are the first k eigen vectors of the variance-covariance matrix between populations

$$(X - \xi 1')W(X - \xi 1')' \tag{2.9}$$

with respect to M^{-1}. In fact the associated eigen values are the squares of the singular values in (2.5).

The concept of canonical variates (coordinates) was introduced in an early paper by the author (Rao (1948)) for graphical representation of taxonomical units characterized by multiple measurements. This was, perhaps, the first attempt to reduce high dimensional data to two or three dimensions using an objective criterion for purposes of graphical displays. Since then, graphical representation of multivariate data for visual examination of clusters, outliers and other structures in data has been an active field of research. Some of the developments are biplots (Gabriel (1971), Gifi (1990), Gower (1993), Greenacre (1993a)) multidimensional scaling (Kruskal and Wish (1978)), correspondence analysis (Benzécri (1992), Greenacre (1984), (1993b)), Chernoff's faces (Chernoff (1993)) and parallel coordinates (Mahalanobis, Mazumdar and Rao (1949), Wegman (1990)). Cavalli-Sfroza (1991) uses canonical coordinates (variables) in interpreting the evolution of human populations.

3. LOSS OF INFORMATION

3.1 Deficiency in the configuration matrix

In terms of the s.v.d. (2.5), the configuration matrix (2.2) of the profiles is

$$C = C_s = W^{-1/2}(\lambda_1^2 V_1 V_1' + \ldots + \lambda_p^2 V_p V_p')W^{-1/2}$$

while that based on the canonical coordinates (2.6) in E_k is

$$C_k = Y'Y = W^{-1/2}(\lambda_1^2 V_1 V_1' + \ldots + \lambda_k^2 V_k V_k')W^{-1/2}$$

and the deficiency is

$$D_1 = C_s - C_k = W^{-1/2}(\lambda_{k+1}^2 V_{k+1} V_{k+1}' + \ldots + \lambda_p^2 V_p V_p')W^{-1/2}.$$

An overall measure of loss of information is the ratio

$$\frac{\text{trace } W^{1/2} D_1 W^{1/2}}{\text{trace } W^{1/2} C_s W^{1/2}} = 1 - \frac{\lambda_1^2 + \ldots + \lambda_k^2}{\lambda^2} \tag{3.1}$$

where $\lambda^2 = \lambda_1^2 + \ldots + \lambda_p^2$. It is customary to compute the ratios

$$\frac{\lambda_1^2}{\lambda^2}, \frac{\lambda_2^2}{\lambda^2}, \ldots \tag{3.2}$$

expressed as percentages to indicate the relative importance of the canonical coordinates in different dimensions.

3.2 Deficiency in the matrix of distances

It is more important to assess the distortions in the inter profile squared distances due to reduction in dimensionality. The $m \times m$ matrix of these squared distances denoted by S can be computed from the configuration matrix C by the formula

$$S = c1' + 1c' - 2C = (d_{ij}^2) \tag{3.3}$$

where c is the vector of the diagonal elements of C. The corresponding matrix in the reduced space is

$$S_{(k)} = c_{(k)}1' + 1c'_{(k)} - 2C_{(k)}$$

so that the matrix

$$D_2 = S - S_{(k)} = d^2_{ij(k)} \tag{3.4}$$

measures the deficiencies in the distances due to reduction of dimensionality. An over all measure of deficiency is the ratio

$$\frac{\Sigma\Sigma w_i w_j d^2_{ij(k)}}{\Sigma\Sigma w_i w_j d^2_{ij}} = \frac{\lambda^2_{k+1} + \ldots + \lambda^2_p}{\lambda^2_1 + \ldots + \lambda^2_p} \tag{3.5}$$

which is the same as (3.1).

3.3 Deficiency in the dispersion matrix

The weighted dispersion matrix between profiles in the original space is

$$B = (X - \xi 1')W(X - \xi 1')' = (b_{ij}) \tag{3.6}$$

while the corresponding matrix in the reduced k-dimensional space is

$$B_{(k)} = M^{-1/2}(\lambda^2_1 U_1 U'_1 + \ldots + \lambda^2_k U_k U'_k)M^{-1/2} = (b_{ij(k)}). \tag{3.7}$$

The proportion of the between profile variance in the i-th variable explained by the first k canonical variates (coordinates) is

$$b_{ii(k)}/b_{ii},\ i = 1, \ldots, p. \tag{3.8}$$

For an interpretation of the canonical coordinates in different dimensions it would be useful to compute the proportion of variance in each variable explained by each of the canonical variates, i.e., to obtain a decomposition of (3.8) in terms of canonical variates. For this purpose, we introduce the matrices

$$E_1 = M^{-1/2}(\lambda_1 U_1 : \ldots : \lambda_p U_p) = (e_{ij}) \tag{3.9}$$

$$E_2 = (e_{ij}/\sqrt{b_{ii}}) = (f_{ij}) \tag{3.10}$$

where b_{ii} is as defined in (3.6). Let $E_{i(k)}$ be the matrix obtained by retaining only the first k columns in E_i for i=1,2. Then it is seen that

$$E_1 E'_1 = B,\ E_{1(k)} E'_{1(k)} = B_{(k)}. \tag{3.11}$$

Let us consider the matrix $E_{1(k)}$ and define what may be called *canonical coordinates for variables* (CCV) in k dimensions as follows.

Table 1. Canonical coordinates for variables

Variable	dim 1	dim 2		dim k
1	e_{11}	e_{12}	$\cdots$	e_{1k}
2	e_{21}	e_{22}	$\cdots$	e_{2k}
.	.	.	$\cdots$	.
p	e_{p_1}	e_{p_2}	$\cdots$	e_{p_k}

If we plot the variables as points in E^k using the row coordinates in different dimensions, then the scalar products of the vectors representing the variables are the elements of $B_{(k)}$, the best k-dimensional approximation to B.

There is some advantage in plotting the variables using the standardized coordinates (f_{ij}) defined in (3.10) as shown in Table 2.

Table 2. Standardized CCV's and the variance explained by each canonical variate

Variable	Standardized coordinates dim 1 ... dim k	Proportion of variance explained dim 1 ... dim k	total
1	$f_{11} \cdots f_{1k}$	$f_{11}^2 \cdots f_{1k}^2$	Σf_{1i}^2
$\vdots$	$\vdots \quad \vdots$	$\vdots \quad \vdots$	$\vdots$
p	$f_{p1} \cdots f_{pk}$	$f_{p1}^2 \cdots f_{pk}^2$	Σf_{pi}^2

The magnitudes in the right hand block of Table 2 indicate the influence of different variables in each dimension (canonical variate) in the reduced space. This may enable us to associate each dimension with certain variables.

We may plot the variables using the CCV's in the same chart as the canonical coordinates for the profiles. It is seen that all variable points lie inside the unit sphere in E^k, and the variables close to the surface of the sphere have greater influence on the canonical variates.

It may also be mentioned that it is the usual practice in a biplot to represent the i-th variable as a directed line using the direction cosines proportional to the i-th row elements in the matrix

$$E_{1(k)} = M^{-1/2}(U_1 : \ldots : U_k) \tag{3.12}$$

in which case the projections of a profile point in these directions are proportional to the approximate coordinates of the profile in the original space (see Gabriel (1971) and Greenacre (1993a)).

Note 3.1. The choices of M and W as inputs in the analysis for canonical coordinates need some discussion. The choice of M is related to the distance measure between profiles appropriate in a given investigation. In taxonomical classification, M is generally chosen as the inverse of the variance-covariance (dispersion) matrix of the measurements on units within taxa leading to Mahalanobis distance (see Mahalanobis (1936), Rao (1945, 1947)). The matrix W is taken to be diagonal with the i-th diagonal element w_i proportional to the number of individuals sampled from the i-th taxa to estimate its profile. For a chosen M, the configuration of the profiles in the reduced space will depend on W, but is likely to be robust provided the w_i's are not widely different. In the study reported in Rao (1948), all the w_i's were chosen as equal although the sample sizes for different populations were different. However, the choice of w_i's as proportional to sample sizes enables us to test hypotheses on goodness of fit of lower dimensional planes to the observed profiles. For details, the reader is referred to Rao (1973, pp.556-560, 1985).

If we desire that the configuration of a subset of profiles to be better preserved in the reduced space than the others, then we have to give bigger weights to those profiles.

Note 3.2 In many situations we have a data matrix X giving the measurements of p variables made on m individuals without any further information to guide us in the choices of the M and W matrices. In such cases, the usual choices of M and W are the unit matrices and the resulting canonical coordinate analysis is the Principal Component Analysis (PCA) introduced by Hotelling. Some characterizations of the principal components and their applications are given in a paper by Rao (1964). It is also the practice to apply PCA on CX, i.e., after a suitable scaling of the measurements. One choice of C is a diagonal matrix with the i-th diagonal element $c_i = s_{ii}^{-1/2}$, where s_{ii} is the i-th diagonal element of the matrix

$$(X - \bar{X}1')(X - \bar{X}1')'.$$

This procedure is equivalent to using the canonical coordinate analysis choosing $M = C$ and $W = I$. Another possibility which has not been considered before is the choice, $c_i = 1/m_i$ where m_i is a measure of location such as the mean or median of the measurements on the i-th variable.

Note 3.3. A more general problem not considered in this paper is as follows. The basic space is somewhat general with a specified nonnegative proximity index between any two points. Given a set of points with the matrix of proximity indices between points, the problem is to transform the points to a low dimensional Euclidean space such that the inequality relationships between proximity indices are maintained to the extent possible in the corresponding Euclidean distances. Such a transformation is achieved through the algorithm for multidimensional scaling as developed by Kruskal and Wish (1978).

4. TWO WAY CONTINGENCY TABLES

We consider dichotomous categorical data with s rows and m columns and n_{ij} observations in the (i,j)-th cell. Define

$$N = (n_{ij}), n_{i.} = \sum_{j=1}^{m} n_{ij},\ n_{.j} = \sum_{i=1}^{s} n_{ij},\ n = \sum_{1}^{s}\sum_{1}^{m} n_{ij}$$

$$R = \text{Diag}\ (n_{1.}/n, \ldots, n_{s.}/n),\ C = \text{Diag}\ (n_{.1}/n, \ldots, n_{.m}/n)$$

$$P = n^{-1}NC^{-1} = \begin{pmatrix} p_{1|1} & \cdots & p_{1|m} \\ \cdot & \cdots & \cdot \\ p_{s|1} & \cdots & p_{s|m} \end{pmatrix}, \quad \text{column profiles,} \tag{4.1}$$

$$Q = n^{-1}R^{-1}N = \begin{pmatrix} q_{1|1} & \cdots & q_{m|1} \\ \cdot & \cdots & \cdot \\ q_{1|s} & \cdots & q_{m|s} \end{pmatrix}, \quad \text{row profiles} \tag{4.2}$$

$$p = R1,\ q = C1.$$

The problem is to represent the column (row) profiles as points in $E^k, k < s$, such that the Euclidean distances between points reflect specified affinities between the corresponding column (row) profiles.

The technique developed for this purpose by Benzécri (1992) is known as correspondence analysis (CA) which can be identified as canonical coordinate analysis. For instance, for representing the column profiles, we need to choose

$$X = P,\ M = R^{-1},\ W = C \tag{4.3}$$

and apply the analysis of Section 1. Thus one finds the s.v.d. of

$$R^{-1/2}(P - p1')C^{1/2} = \lambda_1 U_1 V_1' + \ldots + \lambda_s U_s V_s' \tag{4.4}$$

giving the coordinates for the column profiles in E^k

$$\begin{pmatrix} \lambda_1 V_1' C^{-1/2} \\ \\ \lambda_k V_k' C^{-1/2} \end{pmatrix}. \tag{4.5}$$

Implicit in this analysis is the choice of measure of affinity between the i-th and j-th profiles as the squared distance

$$d_{ij}^2 = \frac{(p_{1|i} - p_{1|j})^2}{p_1} + \ldots + \frac{(p_{s|i} - p_{s|j})^2}{p_s} \tag{4.6}$$

which is the chisquare distance. The squared Euclidean distance between the points representing the i-th and j-th profiles in E^k, the reduced space, is an approximation to (4.6). Thus the clusters we see in the Euclidean representation is based on the affinities measured by the chisquare distance (4.6).

An alternative to the chisquare distance which has some advantages is the Hellinger Distance (HD) between the i-th and j-th column profiles defined by

$$d_{ij}^2 = (\sqrt{p_{1|i}} - \sqrt{p_{1|j}})^2 + \ldots + (\sqrt{p_{s|i}} - \sqrt{p_{s|j}})^2 \tag{4.7}$$

which depends only on the i-th and j-th column profiles. In such a case, the Euclidean distance in the reduced space between the i-th and j-th column profiles is an approximation to (4.7). For the derivation of canonical coordinates of the column profiles (considered as population) we choose

$$X = \begin{pmatrix} \sqrt{p_{1|1}} & \cdots & \sqrt{p_{1|m}} \\ \cdot & \cdots & \cdot \\ \sqrt{p_{s|1}} & \cdots & \sqrt{p_{s|m}} \end{pmatrix}$$

$$M = I, \; W = C = \text{Diag}\,(n_{.1}/n, \ldots, n_{.m}/n)$$

and consider the s.v.d.

$$(X - \xi 1')C^{1/2} = \lambda_1 U_1 V_1' + \ldots + \lambda_s U_s V_s'. \tag{4.8}$$

We may choose $\xi' = (\xi_1, \ldots, \xi_s)$ with

$$\xi_i = \sqrt{p_i} = \sqrt{n_{i.}/n}, \quad \text{or} \tag{4.9}$$

$$= n^{-1}(n_{.1}\sqrt{p_{i|1}} + \ldots + n_{.m}\sqrt{p_{i|m}}). \tag{4.10}$$

The canonical coordinates in E^k for the column profiles choosing ξ as in (4.9) or (4.10) are

$$\begin{pmatrix} \lambda_1 V_1' C^{-1/2} \\ \\ \lambda_k V_k' C^{-1/2} \end{pmatrix}. \tag{4.11}$$

It can be shown that the statistic

$$4n(\lambda_1^2 + \ldots + \lambda_s^2) \tag{4.12}$$

is asymptotically distributed as chisquare on $(s-1)(m-1)$ degrees of freedom to test for independence in the two way contingency table.

The advantages in using Hellinger distance between profiles are the following.

1. The measure depends only on the profiles of the concerned pair. It is not altered when an extended set of profiles is considered.

2. The measure does not depend on the sample sizes on which the profiles are estimated.

3. The choice of ξ as in (4.10) provides a better approximation of the distances in the reduced space. However, if a representation of the row profiles is also needed we take X as the matrix of the square roots of the elements in Q', where Q is the matrix defined in (4.2) and compute the s.v.d. of

$$(Q' - \eta 1')R^{1/2} = \mu_1 A_1 B_1' + \ldots + \mu_m A_m B_m'$$

leading to the canonical coordinates for row profiles

$$\begin{pmatrix} \mu_1 B_1' R^{-1/2} \\ \mu_k B_k' R^{-1/2} \end{pmatrix} . \tag{4.13}$$

4. If we choose ξ as in (4.9), then

$$(X - \xi 1')C^{1/2} = (\sqrt{\frac{n_{ij}}{n}} - \sqrt{\frac{n_{i.}}{n}\frac{n_{.j}}{n}})$$

which is symmetric in i and j. Then, the same s.v.d. as in (4.8) could be used for computing the canonical coordinates

$$\begin{pmatrix} \lambda_1 U_1' R^{-1/2} \\ \lambda_k U_k' R^{-1/2} \end{pmatrix} .$$

for the row profiles, as in the case of CA.

5. AN EXAMPLE

We consider the data (from Greenacre (1993b)) on 796 scientific researchers classified according to their scientific discipline (as populations) and funding category (as variables) as shown in Table 3.

Table 3. Scientific disciplines by research funding categories

				Funding category			
Scientific discipline		a	b	c	d	e	Total
Geology	G	3	19	39	14	10	85

Biochemistry	B_1	1	2	13	1	12	29
Chemistry	C	6	25	49	21	29	130
Zoology	Z	3	15	41	35	26	120
Physics	P	10	22	47	9	26	114
Engineering	E	3	11	25	15	34	88
Microbiology	M_1	1	6	14	5	11	37
Botany	B_2	0	12	34	17	23	86
Statistics	S	2	5	11	4	7	29
Mathematics	M_2	2	11	37	8	20	78
Total		31	128	310	129	198	796

The canonical coordinates for the first three dimensions and percentage of variance explained by each are given in Table 4 for the analyses based on the chisquare distance (correspondence analysis) and the Hellinger distance (alternative). The formulas (4.3)-(4.5) are used for the analysis based on chisquare and the formulas (4.8)-(4.11) for that based on Hellinger distance.

Table 4. Canonical coordinates for the scientific disciplines in the first three dimensions

Subjects	Chisquare Distance			Hellinger distance		
	Dim1	Dim2	Dim3	Dim1	Dim2	Dim3
G	.076401	.302569	-.087749	-.031140	.167408	-.048245
B1	.179892	-.454996	-.151716	-.129374	-.242174	-.077614
C	.037644	.073353	.042371	-.021144	.040433	.028254
Z	-.327365	.102283	.064515	.138850	.045255	.056894
P	.315552	.026997	.108688	-.165340	.010679	.023844
E	-.117495	-.291712	.107330	.049451	-.129906	.082901
M1	.012766	-.109656	-.041435	-.004913	-.052588	-.008439
B2	-.178695	-.038501	-.129055	.151404	-.036559	-.108025
S	.124638	.014162	.107190	-.066639	.011763	.052571
M2	.106751	-.061316	-.175688	-.050307	-.037572	-.078006
% var.	47.20	36.66	13.11	45.87	34.10	16.57

The plots of the scientific disciplines (subjects) using the canonical coordinates based on the chisquare and Hellinger distances are given in Figures 1 and 2 respectively. The coordinates in the third dimension are plotted on a line on the right hand side of the two dimensional plot. This will be of help in visualizing the plot in three dimensions and in interpreting the distances in the two dimensional plot. Thus, although B_2 and E appear to be close to each other in the two dimensional chart, they are clearly separated in the third dimension. No additional distances in the third dimension are involved in the case of P, C, S, Z and E. [It may be noted that the squared distance between any two points in the three dimensional plot is equal to the sum of the squared distances in the two dimensional plot and in the third dimension].

[Here Figures 1 and 2]

It is of interest to note in this example that the configuration of the scientific disciplines in three dimensions obtained by both the methods are very similar. The percentage variance explained in each dimension is nearly the same for both the methods.

The canonical coordinates for the variables (funding categories) are computed using the formulas (3.9)-(3.10). These are obtained from the same s.v.d. used to compute the canonical coordinates for the scientific disciplines. Table 5 gives the standardized canonical coordinates for the funding categories, a, b, c, d, e, using the formula (3.10) for the analyses based on the chisquare and Hellinger distances.

Table 5. Standardized canonical coordinates for funding categories (variables) in the first three dimensions

Funding category	Chisquare Distance				Hellinger Distance			
	dim1	dim2	dim3	%var	dim1	dim2	dim3	%var
a	.758	.114	-.619	97.1	-.796	-.164	-.573	98.9
b	.535	.728	-.137	83.5	-.438	-.766	-.008	77.9
c	.583	.352	.694	94.6	-.501	-.327	.759	93.4
d	-.426	.331	-.172	99.8	.888	-.358	-.285	99.7
e	-.108	-.909	-.081	99.6	.088	.978	-.159	98.9

The standardized canonical coordinates for the funding categories are plotted in Figure 3 (for chisquare distance) and in Figure 4 (for Hellinger distance). It may be noted that all the points lie within the unit circle. It is customary to represent the canonical coordinates for the subjects and variables in one chart. We are using separate charts in order to explain the salient features of the configuration of the variables. The following interpretations emerge from the study of Table 5 and Figures 3 and 4.

[Here Figures 3 and 4]

1. The configurations of the funding categories as exhibited by Figures 3 and 4 obtained by using chisquare and Hellinger distances are very similar.

2. All most all the variation in the funding categories a, d and e is captured in the first three canonical coordinates of the scientific disciplines. A large percentage of variation in b and c is explained by the first three coordinates.

3. The first dimension is strongly influenced by a and d, the second dimension by b and e, and the third dimension by a and c.

Thus the use of standardized coordinates for variables enables us to interpret the different dimensions in terms of observed variables. There are other ways of plotting the coordinates of the variables as mentioned in the paragraphs below Table 3.2. Such biplots having a different interpretation are discussed in Gabriel (1971), Gifi (1990), Gower (1993) and Greenacre (1993a).

6. CONCLUSION

A general theory is developed for representing high dimensional "population by variable" data in a low dimensional Euclidean space. The first step in the problem is the specification of the basic space in which the populations can be considered as points and the choice of a distance measure between points characterizing the differences between populations, which in a practical problem depends on the nature of the investigation involved. The second step is the transformation of the points (representing the populations) from the basic space to a low dimensional Euclidean space. This has to be done in such a way that the configurations of the points as defined by the specified distances in the basic space and the Euclidean distances in the reduced space are as similar as possible. A closed form solution is obtained when the basic space is a vector space endowed with an inner product and the associated norm.

When we have data in the form of a two way contingency table with the frequency in the (i, j)-th cell representing the number of individuals from population j having category i of an attribute, a well known method for representing the populations in a low dimensional Euclidean space is correspondence analysis. The basic space in this case is a vector space where each population is represented by the vector of relative frequencies of the different categories of an attribute and distance between vectors is defined by a chisquare type formula. Such a distance function is not an intrinsic measure of difference between two populations as it depends not only on the differences between their relative frequencies, but also on the average relative frequencies computed from the set of populations under study. Thus the configuration of any subset of populations depends on what other populations are included in the analysis, and also on the relative numbers of individuals observed from each population. An alternative approach of representing a population by the vector of the square roots of relative frequencies and defining distance between two populations by the Hellinger formula does not have the drawbacks associated with the chisquare type formula. In addition, the new analysis has the same advantage of providing tests of significance for homogeneity (or dimensionality) of the populations as in correspondence analysis based on the chisquare formula. Thus the method of dimensionality reduction based on Hellinger distance appears to be a better tool than that on the chisquare distance in exploratory data analysis.

A method of biplots which enables an interpretation of different dimensions in the reduced Euclidean space in terms of the original variables is discussed and illustrated through an example.

The research for this paper is supported by the Army Research Office under Grant DAAH04-96-1-0082.

REFERENCES

1. Benzécri, J.P. (1992). *Correspondence Analysis Handbook.* Marcel Dekkar, Inc., New York.
2. Cavallis-Sfroza, L.L. (1991). Genes, peoples and languages. *Scientific American.* 265, 104-110.
3. Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *J. Amer. Statist. Assoc.* 68, 361-368.
4. Gabriel, K.R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrika* 58, 453-467.
5. Gifi, A. (1990). *Nonlinear Multivariate Analysis.* New York : John Wiley.
6. Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis.* London: Academic Press.
7. Greenacre, M.J. (1993a). Biplots in correspondence analysis. *J. Applied Statistics.* 20, 251-269.
8. Greenacre, M.J. (1993b). *Correspondence Analysis in Practice.* Academic Press, San Diego.
9. Gower, J.C. (1993). Recent advances in biplot methodology. In *Multivariate Analysis: Future Directions 2* (Eds. C.M. Cuadras and C.R. Rao), North Holland, 295-325.
10. Kruskal, J.B. and Wish, M. (1978). *Multidimensional Scaling.* Sage Publications.
11. Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci.* India 12, 49-55.
12. Mahalanobis, P.C., Mazumdar, D.N. and Rao, C.R. (1949). Anthropometric survey of United Provinces, 1941. A statistical study. *Sankhyā* 9, 90-324.
13. Rao, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Cal. Math. Soc.* 37, 81-91.
14. Rao, C.R. (1947). The problem of classification and distance between two populations. *Nature* 159, 30.
15. Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification (with discussion). *J. Roy. Statist. Soc.* Series B10, 159-193.

16. Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā* $\underset{\sim}{26}$, 329-357.

17. Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd Edition, New York : Wiley.

18. Rao, C.R. (1979). Separation theorems for singular values of matrices and their applications in multivariate analysis. *J. Multivariate Analysis* $\underset{\sim}{9}$, 362-377.

19. Rao, C.R. (1980). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In *Multivariate Analysis V* (Ed. P.R. Krishnaiah), Amsterdam: North Holland, 3-22.

20. Rao, C.R. (1985). Tests for dimensionality and interaction of mean vectors under general and reducible covariance structures. *J. Multivariate Analysis* $\underset{\sim}{16}$, 173-184.

21. Von Neumann, J. (1937). Some matrix inequalities and metrization of metric spaces. *Tomsk. Univ. Rev.* $\underset{\sim}{1}$, 286-299.

22. Wegman, E.J. (1990). Hyperdimensional data analysis using parallel coordinates. *J. Amer. Statist. Assoc.* $\underset{\sim}{85}$, 664-675.

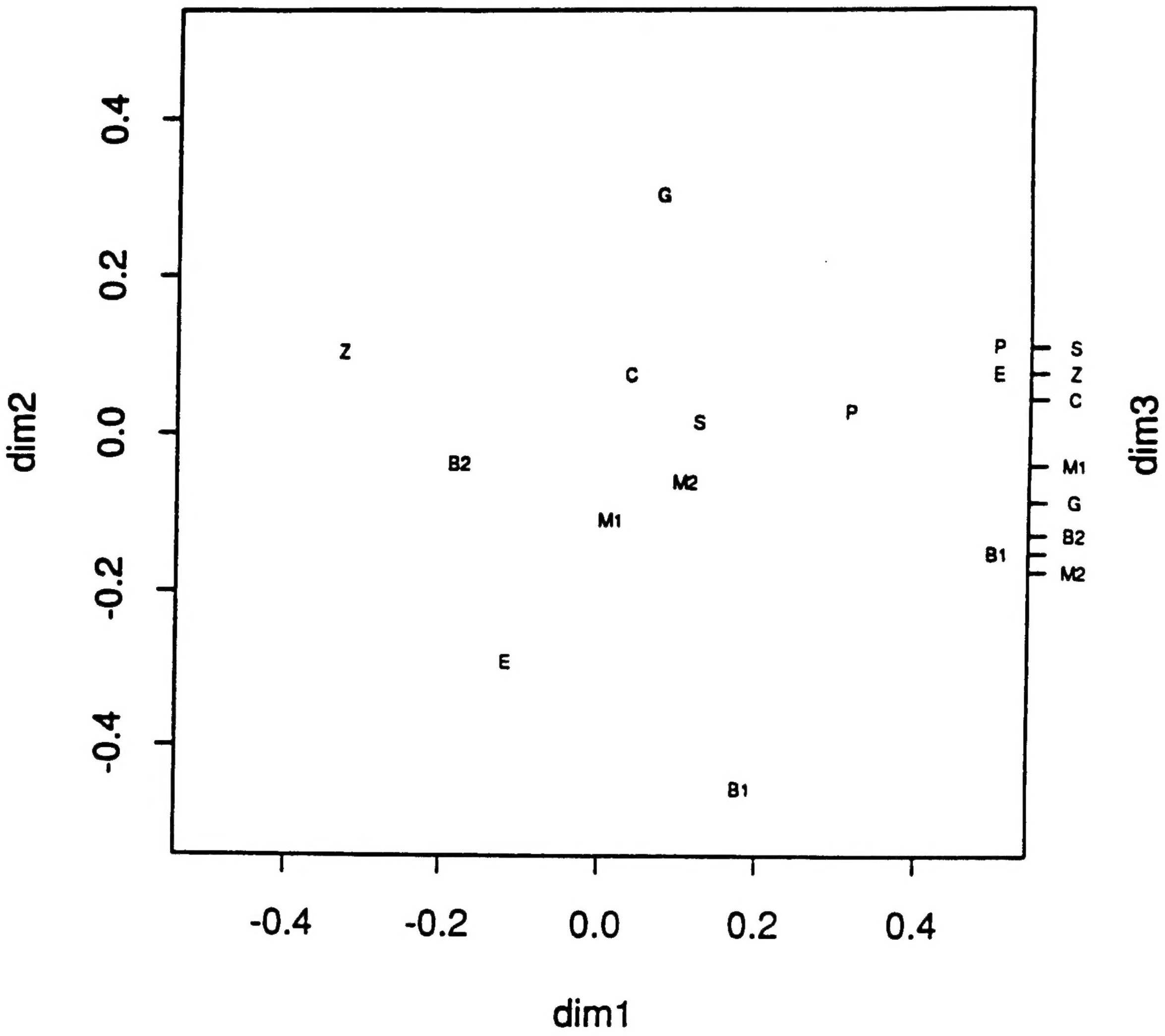

Figure 1. Configuration of scientific disciplines
using Chisquare distance
(Correspondence Analysis)

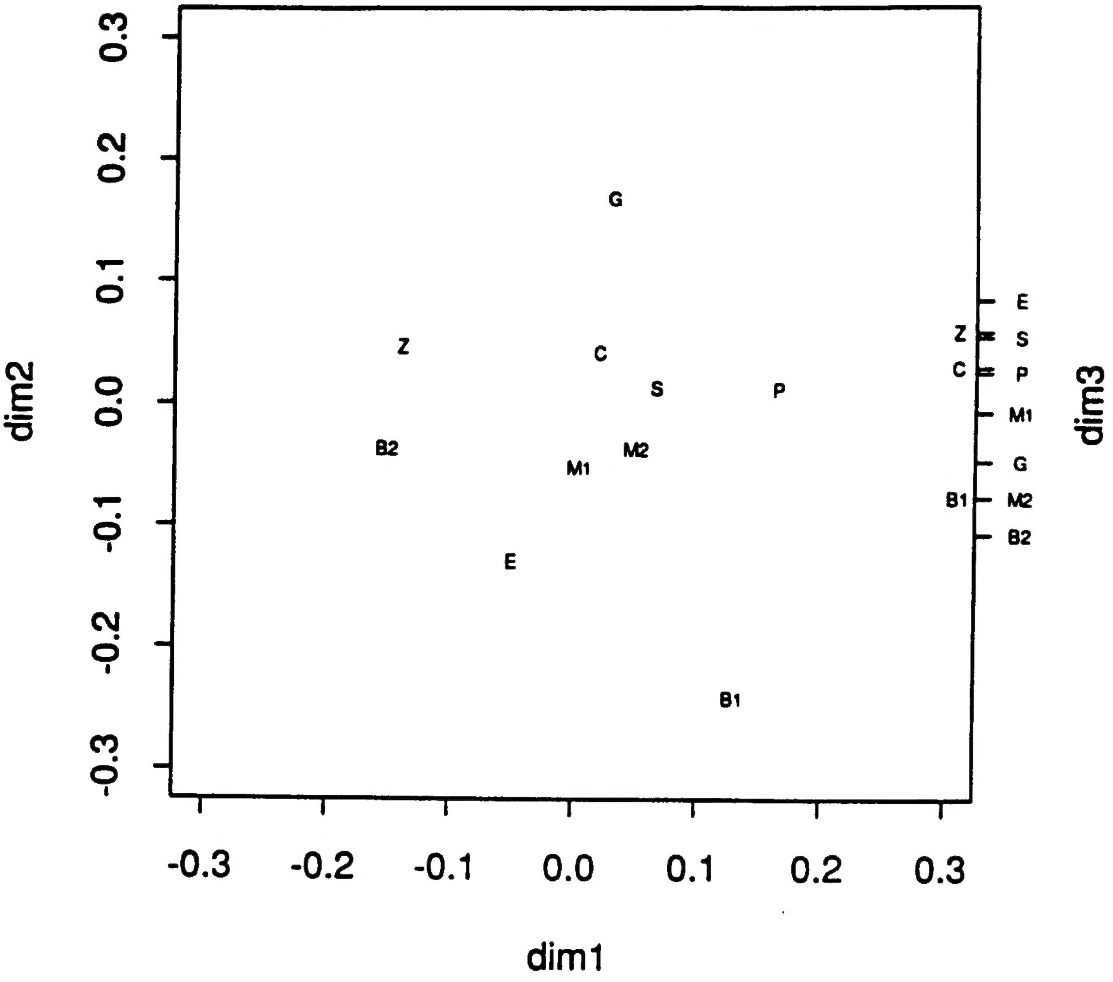

Figure 2 Configuration of scientific disciplines
using Hellinger distance
(Alternative to Correspondence Analysis)

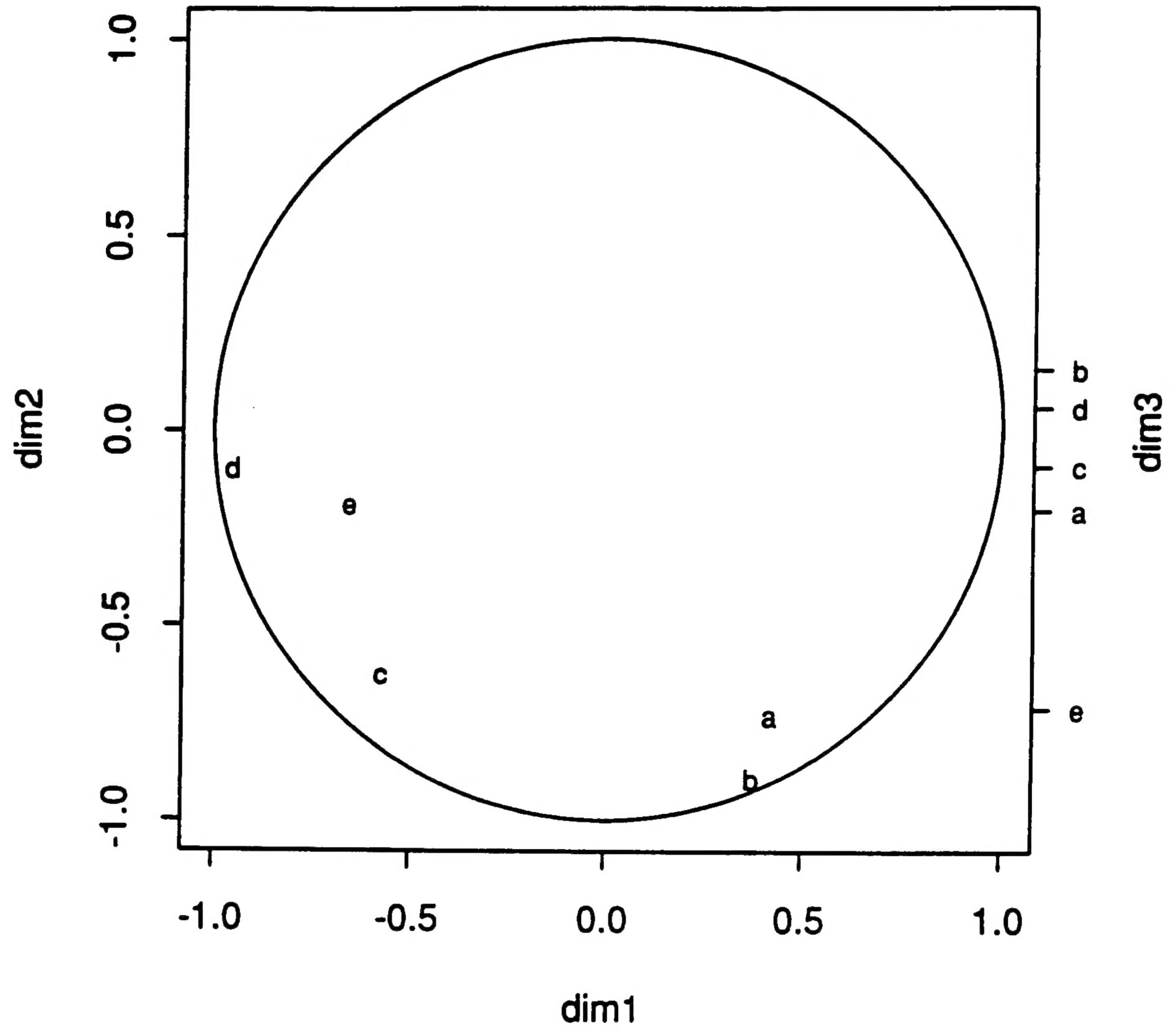

Figure 3. Configuration of funding categories using standardized canonical coordinates based on Chisquare distance

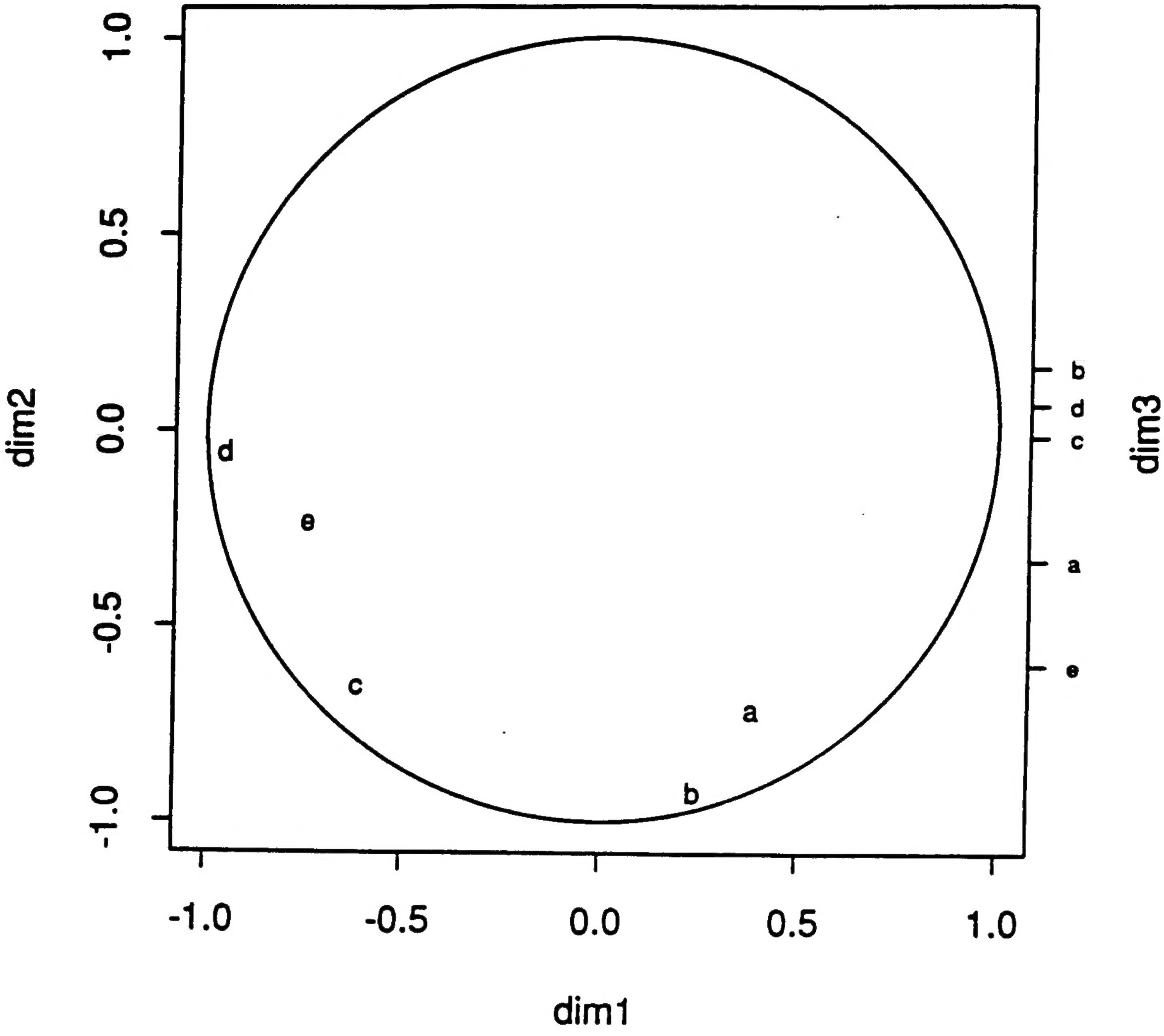

Figure 4. Configuration of funding categories using standardized canonical coordinates based on Hellinger distance